STATISTICS REVISITED: A REVIEW FOR CONTRIBUTORS AND READERS

Size does matter

Jill Mollison, BSc, CStat, *Lecturer in Medical Statistics, Department of Public Health, University of Aberdeen, Aberdeen, UK;* **Julie A Simpson**, PhD, CStat, *Biostatistician, Department of General Practice and Primary Care, University of Aberdeen, Aberdeen, UK;* **Philip C Hannaford**, MD, FRCGP, *Grampian Health Board Chair of Primary Care, Department of General Practice and Primary Care, University of Aberdeen, Aberdeen, UK*

Correspondence: *Jill Mollison, Department of Public Health, University of Aberdeen, Medical School, Polwarth Building, Foresterhill, Aberdeen AB25 2ZD, UK. E-mail: j.mollison@abdn.ac.uk*

(Accepted 30th October 2002)

Journal of Family Planning and Reproductive Health Care 2003: **29(1)**: 9–11

Introduction

When designing an epidemiological study or clinical trial it is important to make it large enough to have a reasonable chance of detecting differences between groups that really exist. In other words, the study should have adequate statistical power. Unfortunately the scientific literature is cluttered with numerous small studies reporting negative results. Individually, each study can only make a modest contribution to clinical practice since it is impossible to know whether a negative result was due to a true lack of effect or limited ability to detect an effect. Absence of evidence is not the same as evidence of absence. In this paper we describe the concepts behind statistical power, including the pieces of information needed when determining the sample size of a study (i.e. how many individuals need to be selected from the study population).

Clinical and statistical significance

Clinical significance relates to effects that are considered to be clinically important whilst statistical significance relates to the likelihood that observed effects could have arisen simply by chance. Studies, particularly large ones, can observe statistically significant differences that are clinically unimportant. Conversely, small studies may observe clinically important effects that are not statistically significant (and so may have been chance findings). Ideally, clinical and statistical significance should match so that real clinically important effects are also statistically significant. In order to optimise the chances of this happening, studies need to perform sample size calculations before starting so that the appropriate number of participants can be recruited.

Information required for a sample size calculation

The information required for a sample size calculation is summarised in Table 1.

Minimum clinical difference between groups (effect of the intervention)

Specifying the minimum *clinically important* difference (i.e. effect size) between groups that the study needs to detect is a key piece of information needed for a sample size calculation. For example, in a trial of two intrauterine contraceptive devices an absolute 10% lowering of the expulsion rate associated with one device (e.g. from 30% to 20%) may be deemed to be clinically worthwhile. In other words, the trial designers deem an absolute 10% difference to be the smallest difference between devices that is clinically significant and therefore worth implementing into clinical practice. In another example, a cohort study may wish to investigate whether use of the oral contraceptive pill (OCP) is associated with an increased risk of myocardial infarction. In order to determine the sample size, the study designers need to specify the minimum clinically important effect (association) that must be detected, perhaps a 50% increased risk among OCP users [i.e. the study needs to be able to detect a relative risk (RR) of 1.5].

Significance level and power (Type I and Type II errors)

In research we are interested in making inferences about the population, usually by studying a sample of the population. We often test the null hypothesis that there is ‘no association’ or ‘no difference’ between groups in a study.[1] When making decisions based on such statistical (hypothesis) testing, two potential errors can occur.

A Type I error *(alpha)* occurs when the null hypothesis is rejected although in fact the null hypothesis is true. This can be considered as a false-positive result as the observed association is interpreted as real when actually it is not. By convention, we often reject the null hypothesis if the statistical test used gives a p value of equal to, or less than, 0.05.[1] This is known as a significance level of 5%. This level of significance means that we incorrectly interpret an association as real on, or less than, 5% of occasions. Setting a higher level of statistical significance ensures that we erroneously interpret an observed association as real on fewer occasions (e.g. if we use the threshold of $p \leq 0.01$ we would be wrong on, or less than, 1% of occasions). Such stringency requires a greater sample size.

Table 1 *Information required for sample size calculation*

Parameter	Description
Minimum clinical difference (effect)	The smallest difference in effectiveness (or association) that is deemed to be clinically significant
Significance level	Typically 5% or 1%
Power	Typically 80% or 90%
Variability of outcome (continuous variables only)	Standard deviation: obtained from previous research or a pilot study
Level of outcome in the baseline group (categorical variables only)	Proportion of patients with the outcome of interest in the baseline group (e.g. standard care, placebo, control or non-exposed group): obtained from previous research or a pilot study
Predicted response and/or loss to follow-up rates	Dependent on type of study and population group (e.g. 20%)

A Type II error *(beta)* occurs when the sample data do not indicate that the null hypothesis should be rejected when, in fact, the null hypothesis is false. This is equivalent to a false-negative result; the study failed to find a statistically significant difference when one really does exist. Typically, the Type II error is set at 0.2 or 0.1. The power of a statistical test is 1-*beta*, e.g. 0.8 or 0.9 (proportions usually expressed as percentages of 80% and 90%, respectively). Statistical power is the likelihood that a study of a given size will detect as statistically significant a real difference between groups of a given magnitude.[2] The key point is to design studies with sufficient power to be reasonably confident of detecting a worthwhile effect or association if it exists. With 80% power a study will be able to detect an important association on 80% of occasions, and will miss it on 20% of occasions. Increasing the power of a study means that important associations are missed on fewer occasions, but at the cost of requiring larger sample sizes.

For the most common situations, further information is required. When the outcome of interest is continuous it is necessary to have information about the variability of the variable in the population of interest. When the outcome is categorical, an estimate of the outcome frequency in the baseline group (e.g. those receiving standard care, placebo or no exposure) is required.

Continuous variable outcomes

When the outcome of intervention or exposure is continuous (e.g. blood loss, blood pressure) a measure of the variability of the outcome measure is required. Therefore, it is important to have an estimate of the standard deviation (SD) of the continuous outcome measure. This can often be obtained from published reports of other studies conducted on a comparable population or from a pilot study.

Categorical variable outcomes

When the outcome is categorical, the expected proportion of subjects experiencing the outcome of interest in the baseline group is required. This information can also often be obtained from published reports or a pilot study.

Predicted response rates and loss to follow-up

Many studies collect data through patient questionnaires and/or follow up participants over a period of time. Not all individuals initially invited to participate in a study will provide data and some initial recruits will be lost to follow-up. When performing sample size calculations it is worthwhile increasing the sample size to allow for these anticipated losses. For example, the sample size would need to be inflated by a factor of 1.25 to account for an anticipated 20% non-response rate. The following examples illustrate two common scenarios.

Table 2 *Illustration of variation in sample size for a continuous outcome variable*

	5% Significance level		1% Significance level	
	Minimum clinical difference		Minimum clinical difference	
Power	5 days	3 days	5 days	3 days
80%	64	176	96	262
90%	86	235	121	333

Numbers in table denote number of subjects required in *each* group.

Example 1: Comparison of continuous outcome between two independent groups

A study is to be conducted to evaluate the effect of mefenamic acid versus placebo for controlling irregular bleeding following Norplant use. A randomised controlled trial (RCT) would be the most appropriate design to address this question, with two groups of women randomly allocated to mefenamic acid or placebo. One of the main outcome measures is to be the number of days of irregular bleeding. For the purpose of the sample size calculation we will assume that number of days of irregular bleeding is Normally distributed. Information needed for the sample size calculation includes the minimum clinically important difference between groups and the variability of number of days of irregular bleeding in Norplant users. A previously published paper reported that the SD of the number of days of irregular bleeding was 10 days.[3] Taking a difference of 5 days as the minimum clinically important difference, we need to calculate the number of patients required for a trial that wished to detect a standardised difference (minimum clinically important difference/SD) of 0.5. To ensure 80% power of detecting a difference of at least 5 days, at the 5% significance level, the trial would need to recruit 64 women in each group.[4] However, if a difference of 3 days were considered to be clinically significant, the required sample size would increase to 176 women per group; larger sample sizes being required to detect smaller differences. Table 2 shows how the sample size increases as the power and significance levels become more stringent.

Example 2: Comparison of categorical outcome between two independent groups

A cohort study is to be conducted to assess the association between OCP use and headache. In cohort studies the size of association is often considered in terms of a relative risk (RR) rather than absolute differences between groups. To calculate the required sample size for this cohort study, an estimate of the frequency of headache in non-OCP users is required (perhaps 10%). In addition, an estimate of the minimum size of effect deemed to be clinically important is required (perhaps a RR of 1.5). A RR of 1.5 applied to the baseline frequency in non-OCP users gives a frequency of headaches in OCP users of 15%, and an absolute difference between groups of 5% (15–10%). Table 3 shows that for the study to have 80% power of detecting at the 5% significance level an absolute difference between groups of 5% (i.e. RR = 1.5), 726 women would have to be recruited into both the OCP user and non-user groups.[4] If the frequency of headache in the baseline group was higher (e.g. 20%), smaller sample sizes would be required to detect the same RR. This is because the absolute differences between groups is larger (a RR of 1.5 applied to a baseline frequency of headaches in non-OCP users of 20% produces a 30% frequency of headaches in OCP users and an absolute difference between groups of 10%). Table 3 also shows that study sample sizes increase as associations become smaller, and as the power of a study increases.

Some additional comments

Estimating sample sizes is not an exact science. Pragmatic compromises often have to be struck between having enough individuals in a study to provide reasonable if not necessarily optimal statistical power while at the same time ensuring that there are sufficient resources to complete the task. Partly for these reasons, many studies use 80% power and 5% significance when making their sample size calculations.

Table 3 *Illustration of variation in sample size for a categorical outcome variable*

	Level of outcome in control cohort					
	10%			20%		
	Minimum clinical difference			Minimum clinical difference		
	RR = 1.2	RR = 1.5	RR = 2.0	RR = 1.2	RR = 1.5	RR = 2.0
Power	(10–12%)	(10–15%)	(10–20%)	(20–24%)	(20–30%)	(20–40%)
80%	3941	726	219	1733	313	91
90%	5242	957	286	2302	412	119

Numbers in table denote number of subjects required in *each* cohort at the 5% significance level.
RR, Relative risk.

Often studies fail to identify statistically significant findings. As we have seen this does not necessarily mean that in the population there is no difference in effectiveness between the treatments or no association between an exposure and outcome; the study may simply have lacked statistical power. This possibility needs to be remembered when interpreting the results of negative studies, especially those that report large 95% confidence intervals around the effect size estimates. Researchers are increasingly expected to report the power of their study when publishing their results. Although underpowered studies are undesirable, they sometimes occur because of circumstances beyond the researchers' control, for instance because of unexpected problems with recruitment or follow-up. Individually, underpowered studies should have only limited clinical impact. However, statistical methods for aggregating results from different studies through meta-analysis now mean that collectively underpowered studies can make important contributions to clinical practice. All researchers, therefore, should publish their results even if their study lacks statistical power.

Often in family planning and reproductive health care the aim is to show that one treatment is as effective as another (e.g. different OCP formulations). In this situation the aim is to show that the different OCPs are equivalent, for instance with respect to efficacy. Trials such as these are known as equivalence trials and the sample size should reflect this design.[5] In general, equivalence trials require a greater sample size than trials designed to show that one treatment is more effective than another (i.e. superiority trials).

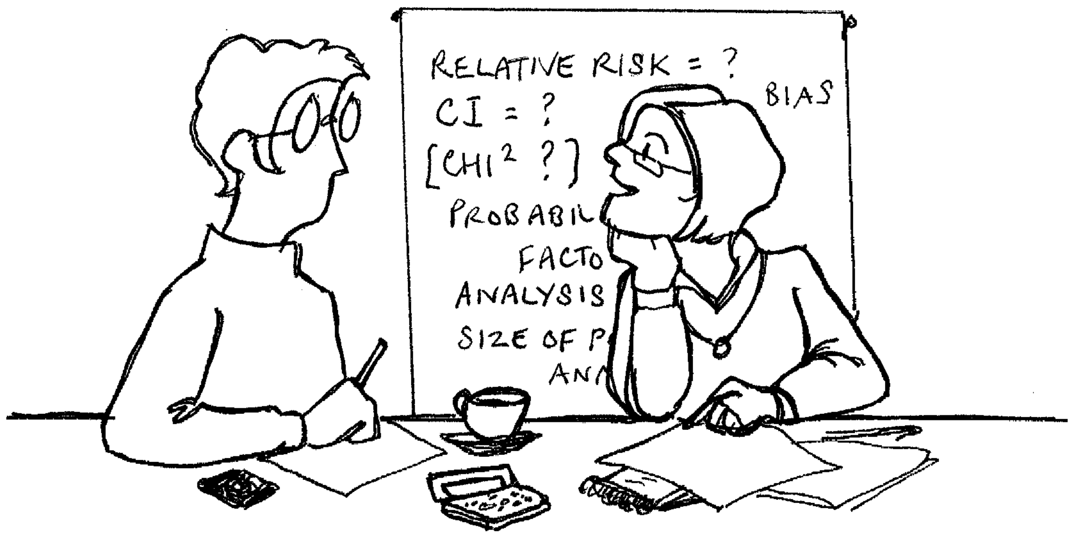
The actual calculations used for obtaining sample size will depend on the study design (e.g. RCT, case-control study), the statistical test to be adopted (e.g. independent groups *t*-test, paired *t*-test, multiple regression) and the type of outcome variable (e.g. continuous, time to an event, ordinal). We have not been able to describe all of these situations in this paper. There are numerous books,[4] commercial[6] and freeware[7] software packages that will compute sample sizes. For the comparison of a continuous variable a simple nomogram exists for estimating the required sample size.[8] It is likely, however, that most readers will wish to consult a statistician when designing their study in order to get help with the sample size calculation. This is important for ensuring that the study is well designed, for gaining funding and for optimising the chances of publication of the study findings. When consulting the statistician, the encounter is likely to be of greater mutual benefit if the researcher is able to provide from the start the key pieces of information needed for the sample size calculation.

Statements on funding and competing interests
Funding. None identified.
Competing interests. None identified.

References

1 Mollison J, Simpson JA, Hannaford PC. From samples to populations: estimation and hypothesis testing. *J Fam Plann Reprod Health Care* 2002; **28**: 101–104.
2 Altman DG. *Practical statistics for medical research*. London: Chapman & Hall, 1991.
3 Kaewrudee S, Taneepanichskul S, Jaisamraun U, et al. The effect of mefenamic acid on controlling irregular uterine bleeding secondary to Norplant use. *Contraception* 1999; **60**: 25–30.
4 Machin D, Campbell M, Fayers P, et al. *Sample size tables for clinical studies*. Oxford: Blackwell, 1997 (includes computer program on disk).
5 Piaggio G, Pinol APY. Use of the equivalence approach in reproductive health clinical trials. *Stat Med* 2001; **20**: 3571–3587.
6 nQuery Advisor®. Cork: Statistical Solutions Limited. www.statsol.ie/nquery
7 Epi Info. www.cdc.gov/epiinfo
8 Altman DG. How large a sample? In: Gore SM, Altman DG (eds), *Statistics in practice*. London: British Medical Association, 1982.

Cartoon by Jeanette Cayley



'Do you think that there is a need for a *Journal of Inspired Guesses*?'